Partner 1
Partner 2
Partner 3

CSCI 1550 / 2540
February 8th, 2023

Homework 1

Due: February 29th, 2023
Remember to show your work for each problem to receive full credit.

Problem 1 [50 points]

Consider the following algorithm for finding the k-smallest element in a set S:
Procedure Select(S, k);
Input: A set S, an integer $k \leq |S| = n$.
Output: The k smallest element in the set S.

1. If $|S| \leq 24$ sort S and return the k smallest element. STOP.
2. Choose a random element y uniformly from S.
3. Compare all elements of S to y. Let $S_1 = \{x \in S \mid x \leq y\}$ and $S_2 = \{x \in S \mid x > y\}$.
4. If $k \leq |S_1|$ return Select(S_1, k) else return Select$(S_2, k - |S_1|)$.

Answer the following questions for $|S| = n$ (you can ignore the cost of step 1 which is $O(1)$):

1. We say that a call to Order(S, k) was *successful* if both $|S_1| \leq 2n/3$ and $|S_2| \leq 2n/3$. Prove that the algorithm terminates after no more than $\log_{3/2} n$ successful calls.
2. Prove that a call to the algorithm if $|S| = n \geq 24$ is successful with probability $\geq 1/4$. [Hint: $2n/3$ may not be an integer. S is a set.]
3. Let Y_i be a geometric random variable with parameter $p = 1/4$. Argue (formally or informally) that for the analysis of the algorithm's runtime we can use Y_i as an upper bound on the number of calls between the i-th successful call (excluded) and the $i+1$-th successful call (included).

   We continue the analysis assuming that for all i, the number of calls between the i-th successful call (excluded) and the $i+1$-th successful call (included) is distributed according to Y_i.
4. Let X_i be the number of comparisons between the i-th successful call (excluded) and the $i+1$-th (inluded). Argue that for the analysis of the algorithm's performance, X_i is bounded by $n(2/3)^i Y_i$.

   We continue the analysis assuming that $X_i = n(2/3)^i Y_i$. Under this assumption, prove that $E[X_i] = n(2/3)^i E[Y_i] = 4n(2/3)^i$.
5. Let X be the total number of comparisions executed by the algorithm. Prove that $E[X]$ is bounded by $12n$.

6. Derive $Var[Y_i]$ and $Var[X_i]$.

7. Prove that $Var[X] \leq \sum_{i=0}^{\log_{3/2} n} n^2(2/3)^{2i} Var[Y_i] \leq 21.6n^2$

8. Apply Chebyshev's inequality to prove that with probability ≥ 0.85 the algorithm executes no more than $24n$ comparisons.

Problem 2 [20 points]

Let $a_1, a_2, ..., a_n$ be a list of n distinct numbers. We say that a_i and a_j are inverted if $i < j$ but $a_i > a_j$. The *Bubblesort* sorting algorithm swaps pairwise adjacent inverted numbers in the list until there are no more inversions, so the list is in sorted order. Suppose that the input to *Bubblesort* is a random permutation, equally likely to be any of the $n!$ permutations of n distinct numbers.

a. Determine the expected number of inversions that need to be corrected by *Bubblesort*.

b. Determine the variance of the number of inversions that need to be corrected by *Bubblesort*.

Problem 3 [15 points]

Suppose that we have an algorithm that takes as input a string of n bits. We are told that if the input bits are chosen independently and uniformly at random, the expected running time is $O(n^2)$. What can Markov's inequality tell us about the worst-case running time of this algorithm on inputs of size n? [Hint: What is the sample space? What is the smallest probability of any event in that sample space?]

Problem 4 [15 points]

We have a standard six-sided die. Let X be the number of times that a 6 occurs over n throws of the die. Let p be the probability of the event $X \geq n/4$. Compare the best upper bounds on p that you can obtain using Markov's inequality, Chebyshev's inequality, and Chernoff bounds.

Problem 5 [25 points]

Suppose that we can obtain independent samples $X_1, X_2, ...$ of a random variable X and that we want to use these samples to estimate $\mathbb{E}[X]$. Using t samples, we use $\frac{1}{t}\sum_{i=1}^{t} X_i$ for our estimate of $\mathbb{E}[X]$. We want the estimate to be within $\varepsilon\mathbb{E}[X]$ from the true value of $\mathbb{E}[X]$ with probability at least $1-\delta$. We may not be able to use Chernoff's bound directly to bound how good is our estimate is if X is not a 0-1 random variable, and we do not know its moment generating function. We develop an alternative approach that requires only having a bound on the variance of X. Let $r = \frac{\sqrt{Var(X)}}{\mathbb{E}[X]}$.

a. Show using Chebyshev's inequality that $O(\frac{r^2}{\varepsilon^2\delta})$ samples are sufficient to solve the problem.

b. Suppose that we only need a weak estimate that is within $\varepsilon\mathbb{E}[X]$ of $\mathbb{E}[X]$ with probability at least $\frac{3}{4}$. Show that $O(\frac{r^2}{\varepsilon^2})$ are enough for this weak estimate.

c. Show that by taking the median of $O(\log(\frac{1}{\delta}))$ weak estimates, we can obtain an estimate within $\varepsilon\mathbb{E}[X]$ of $\mathbb{E}[X]$ with probability at least $1-\delta$. Conclude that we need only $O(\frac{r^2\log(\frac{1}{\delta})}{\varepsilon^2})$ samples.

   Hint: Let Y_i be the i^{th} weak estimate and let Y be the median of all the weak estimates. Show that $|Y - \mathbb{E}[X]| \geq \varepsilon\mathbb{E}[X]$ implies that at least half of the Y_i's satisfy $|Y_i - \mathbb{E}[X]| \geq \epsilon\mathbb{E}[X]$).